Лекция по эконометрике № 5

Множественная линейная регрессия

Демидова
Ольга Анатольевна
https://www.hse.ru/staff/demidova_olga
E-mail:demidova@hse.ru
05.10.2020

План лекции № 5

- Прогнозирование по модели парной регрессии
- Доверительные интервалы для среднего и индивидуального прогноза
- Проверка нормальности распределения
- Множественная линейная регрессия в скалярной и матричной формах
- Метод наименьших квадратов и его геометрическая интерпретация в многомерном случае. Система нормальных уравнений
- Матричное выражение для вектора оценок коэффициентов регрессии

Прогнозирование по модели парной регрессии

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1,\ldots, n$$

Прогноз для $\mathrm{X}_{\mathrm{n}+1}$ – ?

$$Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}$$

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}$$

Ошибка индивидуал ьного прогноза

$$\mathrm{e}_{\mathrm{n}+1} = Y_{n+1} - \hat{Y}_{n+1} = (\beta_0 - \hat{\beta}_0) + X_{n+1}(\beta_1 - \hat{\beta}_1) + \varepsilon_{n+1}$$

$$\mathrm{var}(\mathrm{e}_{\mathrm{n}+1}) = \mathrm{var}(\hat{\beta}_0) + X_{n+1}^2 \, \mathrm{var}(\hat{\beta}_1) +$$

$$+ 2X_{n+1} \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) + \mathrm{var}(\varepsilon_{n+1}) =$$

$$= \sigma_\varepsilon^2 \left[\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2} \right) + \frac{X_{n+1}^2}{\sum_{i=1}^{n} x_i^2} - 2\frac{X_{n+1} \cdot \overline{X}}{\sum_{i=1}^{n} x_i^2} + 1 \right]$$

Прогнозирование по модели парной регрессии

$$\operatorname{var}(\mathrm{e}_{n+1}) = \sigma_\varepsilon^2 \left[\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2} \right) + \frac{X_{n+1}^2}{\sum_{i=1}^{n} x_i^2} - 2\frac{X_{n+1} \cdot \overline{X}}{\sum_{i=1}^{n} x_i^2} + 1 \right]$$

$$= \sigma_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{\sum_{i=1}^{n} x_i^2} \right]$$

Прогнозирование по модели парной регрессии

$$\frac{Y_{n+1} - \hat{Y}_{n+1}}{\sqrt{\sigma_{\varepsilon}^{2}\left[1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X})^{2}}{\sum_{i=1}^{n} x_i^2}\right]}} \sim N(0,1)$$

$$\frac{Y_{n+1} - \hat{Y}_{n+1}}{\sqrt{\hat{\sigma}_{\varepsilon}^{2}\left[1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X})^{2}}{\sum_{i=1}^{n} x_i^2}\right]}} \sim t(n-2)$$

Прогнозирование по модели парной регрессии

Доверительный интервал для индивидуального прогноза

$$\hat{\beta}_0 + \hat{\beta}_1 X_{n+1} \pm t_{\alpha/2} \sqrt{\hat{\sigma}_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{\sum_{i=1}^{n} x_i^2} \right]},$$

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-2}$$

Прогнозирование по модели парной регрессии

$$Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1} \text{ – индивидуальный прогноз}$$

$$E(Y_{n+1}) = \beta_0 + \beta_1 X_{n+1}$$

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1} \text{ – "средний" прогноз}$$

Ошибка "среднего" прогноза

$$\tilde{e}_{n+1} = E(Y_{n+1}) - \hat{Y}_{n+1} = (\beta_0 - \hat{\beta}_0) + X_{n+1}(\beta_1 - \hat{\beta}_1) =$$

$$\operatorname{var}(\tilde{e}_{n+1}) = \operatorname{var}(\hat{\beta}_0) + X_{n+1}^2 \operatorname{var}(\hat{\beta}_1) +$$

$$+ 2 X_{n+1} \operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) =$$

$$= \sigma_\varepsilon^2 \left[\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2} \right) + \frac{X_{n+1}^2}{\sum_{i=1}^{n} x_i^2} - 2 \frac{X_{n+1} \cdot \overline{X}}{\sum_{i=1}^{n} x_i^2} \right]$$

Прогнозирование по модели парной регрессии

Доверительный интервал для "среднего" прогноза

$$\hat{\beta}_0 + \hat{\beta}_1 X_{n+1} \pm t_{\alpha/2} \sqrt{\hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{\sum_{i=1}^{n} x_i^2} \right]},$$

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-2}$$

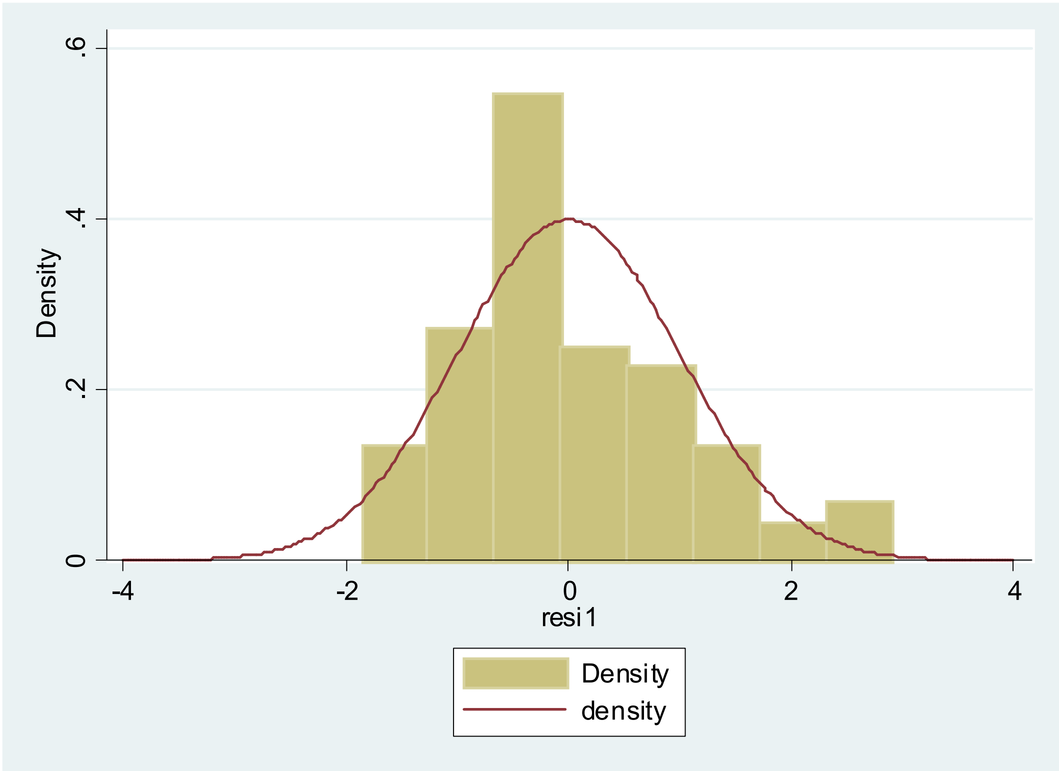
Тестирование регрессионных остатков на нормальность распределения

9

Проверка нормальности распределения остатков

Визуальный анализ

•**Сравнение гистограммы остатков с гистограммой нормального распределения**

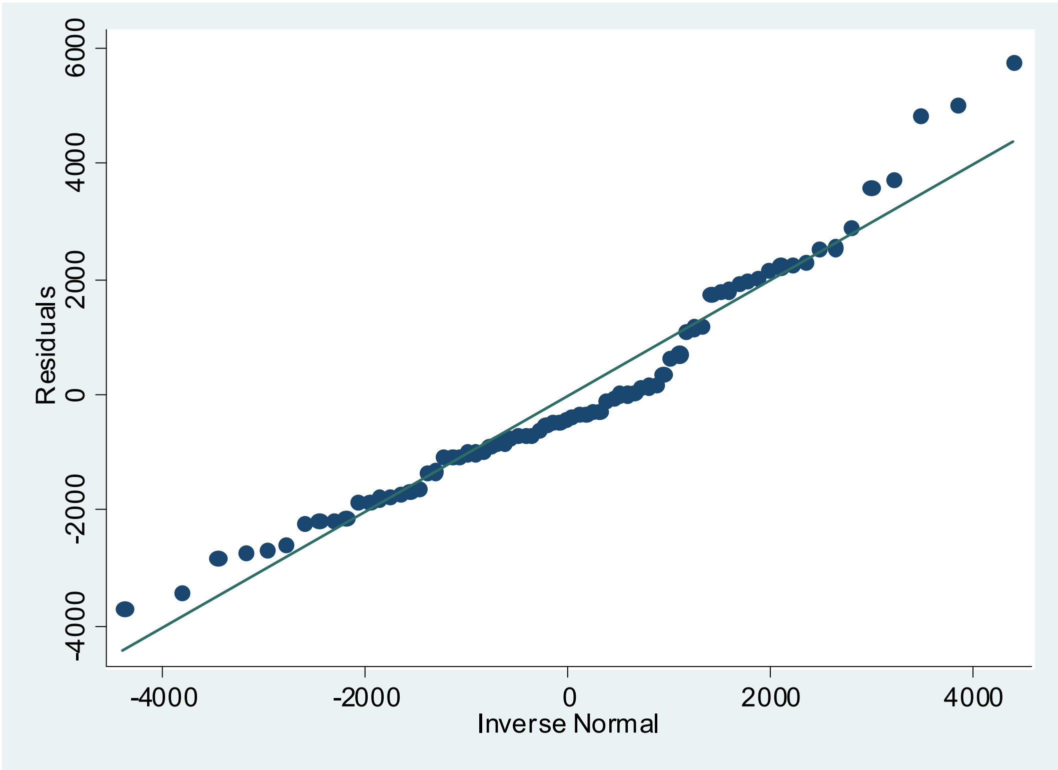


10

Проверка нормальности распределения остатков

Визуальный анализ

Q-Q plot (Q-norm plot)



Проверка нормальности распределения остатков

Тест Jarque-Bera

$$H_0 : e_i \sim N(.,\ .)$$

$$H_1 : e_i \not\subset N(.,.)$$

$$JB = \frac{n}{6}\left(sk^2 + \frac{1}{4}(k-3)^2 \right) \sim \chi^2(2)$$

Sk – skewness, k – kurtosis (нормированные третий и четвертый центральные моменты)

$$sk = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{(\mathrm{X}_{\mathrm{i}} - \overline{X})^3}{\hat{\sigma}^3}; \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}} (X_i - \overline{X})^2;$$

$$k = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{(\mathrm{X}_{\mathrm{i}} - \overline{X})^4}{\hat{\sigma}^4}$$

Проверка нормальности распределения остатков

Недостаток: Тест Jarque-Bera применим только при большом числе наблюдений, при малом следует использовать тест Шапиро – Уилка.

Весьма популярным является тест Колмогорова-Смирнова проверки нормальности.

В этих тестах основная гипотеза состоит в том, что остатки имеют нормальное распределение.

Если p-value < α (выбранный уровень), то основная гипотеза отвергается.

Множественная линейная регрессия

$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$, $i = 1, \ldots, n$ -

общий вид модели множественной регрессии.

$X_1, \ldots, X_k$ – факторы (независимые переменные),

Y – зависимая переменная,

ε_i – возмущения,

n – число наблюдений.

Множественная линейная регрессия

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i, \; i = 1, \ldots, n$$

Обозначим

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}, X_1 = \begin{pmatrix} X_{11} \\ X_{12} \\ \ldots \\ X_{1n} \end{pmatrix}, \ldots, X_k = \begin{pmatrix} X_{k1} \\ X_{k2} \\ \ldots \\ X_{kn} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix}$$

Тогда уравнение регрессии можно переписать в векторном виде

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \ldots + \beta_k \mathrm{X}_k + \varepsilon$$

Множественная линейная регрессия

Если ввести матрицу наблюдений X размера (nx(k+1)) и

вектор коэффициентов β размера ((k+1)x1)

$$X = \begin{pmatrix} 1 & X_{11} & \dots & X_{k1} \\ 1 & X_{12} & \dots & X_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & X_{1n} & \dots & X_{kn} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix},$$

то уравнение регрессии можно переписать в матричном виде:

$$Y = X\beta + \varepsilon$$

Оцененные значения зависимой переменной и остатки регрессии

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon,$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki}, i = 1, \ldots, n$$

$$e_i = Y_i - \hat{Y}_i, i = 1, \ldots, n$$

МНК для множественной линейной регрессии

$$Y = X\beta + \varepsilon$$

$$\hat{Y} = X\hat{\beta}$$

$$e = Y - \hat{Y} = Y - X\hat{\beta}$$

МНК для множественной линейной регрессии

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$\sum_{i=1}^{n} e_i^2 = (e_1, \dots, e_n) \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = e'e$$

МНК для множественной линейной регрессии

$$RSS(\hat{\beta}) = e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta}) =$$

$$= (Y' - \hat{\beta}'X')(Y - X\beta) =$$

$$= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$\underset{(1\times n)}{Y'}\ \underset{(n\times k)}{X}\ \underset{(k\times 1)}{\hat{\beta}} \quad \Rightarrow (Y'X\hat{\beta}) = (Y'X\hat{\beta})' = \hat{\beta}'X'Y$$

$$RSS(\hat{\beta}) = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

Необходимые условия экстремума, система нормальных уравнений и оценка МНК

$$RSS(\hat{\beta}) = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$\frac{\partial RSS(\hat{\beta})}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Достаточное условие экстремума

$$RSS(\hat{\beta}) = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$H = \frac{\partial^2 RSS(\hat{\beta})}{\partial\hat{\beta}\partial\hat{\beta}'} = 2X'X$$